Nonlinear Utility and the Probability Score[1,2]

Robert L. Winkler

Graduate School of Business, Indiana University, Bloomington

and Allan H. Murphy[3]

Traverlers Research Corporation, Hartford, Conn.

(Manuscript received 26 August 1969, in revised form 27 October 1969)

ABSTRACT

Proper scoring rules, such as the probability score, are based (in part) upon the assumption that the *assessor's* utility function is linearly related to the score. The effects of two nonlinear utility functions, one representing a "risk-taker" and one representing a "risk-avoider," on an assessor's probability forecasts are considered. The results indicate that factors other than the expected score, e.g., the variance of the score, may be relevant for probability assessment. In general, a "risk-taker" "hedges" toward a categorical forecast, while a "risk-avoider" "hedges" away from a categorical forecast. The implications of these results for the process of probability assessment are briefly discussed.

1. Introduction

The probability score, or Brier score (Brier, 1950), is considered by many meteorologists to be the "best" available measure of the "goodness" of probability forecasts (Murphy, 1969). Of particular interest in this paper is the fact that the probability score is a member of the class of "proper scoring rules" (Winkler and Murphy, 1968). A proper scoring rule is a scoring rule which forces the assessor (the meteorologist) to make his statements (forecasts) correspond to his subjective probabilities (judgments or beliefs) in order to maximize or minimize his expected score (with some scoring rules a higher score is "better," while with other scoring rules a lower score is "better"). Since in the case of the probability score a lower score is "better," the assessor should attempt to minimize his expected score (the expectation being taken with respect to his subjective probabilities).

In the development of proper scoring rules we (and others) have assumed that the assessor attaches a utility to each possible score and that this utility is linearly related to the score. Alternatively, the assessor could be given a monetary payoff which is linearly related to the score, in which case the postulates of decision theory (Savage, 1954; Fishburn, 1964) imply that he should attempt to maximize his expected payoff. The maximization of the expected payoff is equivalent to the maximization (or minimization) of the expected score only if the assessor's utility function for money is linear within the relevant range (i.e., within the range of potential payoffs). Thus, if the assessor's utility function for money is nonlinear, factors other than the expected score may affect his probability forecasts.[4] In this paper the effects of two nonlinear utility functions, one representing a "risk-taker" and one representing a "risk-avoider," on an assessor's probability forecasts are considered.[5] Section 2 contains a brief discussion of the probability score. Sections 3 and 4 consider the probability forecasts of the "risk-taker" and the "risk-avoider," respectively. The implications of nonlinear utility functions for probability assessment are considered in Section 5, and Section 6 contains a brief summary and conclusion.

2. The probability score

Consider an individual (the assessor) who must make a probability forecast for an event E which consists of n mutually exclusive and collectively exhaustive outcomes, $E_1, E_2, \cdots, E_n$. Let r_i denote the assessor's probability forecast for the outcome E_i, and let p_i denote the assessor's true judgment regarding the probability that E_i will occur. That is, p_i is the assessor's subjective probability that E_i will occur and r_i is his forecast, or stated probability, that E_i will occur. Of course, $p_i \geq 0$ and $r_i \geq 0$ for $i=1, 2, \cdots, n$, and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} r_i = 1.$$

[1] Supported in part by the Atmospheric Sciences Section, National Science Foundation, under Grant GA-1707.

[2] Contribution No. 165 from the Department of Meteorology and Oceanography, University of Michigan.

[3] Present affiliation: Department of Meteorology and Oceanography, University of Michigan, Ann Arbor.

[4] Note that we are concerned with the *assessor's* (and not the decision maker's) utility function.

[5] For an introduction to the effect of nonlinear utility functions on probability assessment when the scoring rule of concern is the probability score, refer to Murphy (1970).

Further, let d_i equal one if E_i occurs and zero otherwise. The probability score for an individual forecast is then

$$PS=\sum_{i=1}^{n}(r_i-d_i)^2. \quad (2.1)$$

The "best" possible score is zero (when a forecast is categorical and correct, i.e., when $r_i=1$ and $d_i=1$ for some i), and the "worst" possible score is two (when a forecast is categorical and incorrect, i.e., when $r_i=1$ and $d_i=0$ for some i).

If E_h occurs, then PS can be written in the form

$$PS_h=1-2r_h+\sum_{i=1}^{n}r_i^2. \quad (2.2)$$

The expected value of PS, or the expected probability score, is

$$\begin{aligned}E(PS)&=\sum_{h=1}^{n}p_hPS_h\\&=\sum_{h=1}^{n}p_h\left(1-2r_h+\sum_{i=1}^{n}r_i^2\right)\\&=1-2\sum_{h=1}^{n}p_hr_h+\sum_{i=1}^{n}r_i^2\\&=1-2\sum_{i=1}^{n}p_ir_i+\sum_{i=1}^{n}r_i^2. \quad (2.3)\end{aligned}$$

For a given set of subjective probabilities, $p_1, p_2, \cdots, p_n$, $E(PS)$ is minimized when $r_i=p_i$ for $i=1, 2, \cdots, n$. In other words, $E(PS)$ is minimized when the assessor's probability forecast for each E_i coincides with his true judgments about the probability of occurrence of the E_i.

The second moment of the probability score is

$$E(PS^2)=\sum_{h=1}^{n}p_h\left(1-2r_h+\sum_{i=1}^{n}r_i^2\right)^2,$$

which simplifies to

$$E(PS^2)=1+4\sum_{i=1}^{n}p_ir_i^2+\left(\sum_{i=1}^{n}r_i^2\right)^2-4\sum_{i=1}^{n}r_ip_i+2\sum_{i=1}^{n}r_i^2-4\left(\sum_{i=1}^{n}p_ir_i\right)\left(\sum_{i=1}^{n}r_i^2\right).$$

The variance of the probability score is

$$V(PS)=E(PS^2)-[E(PS)]^2, \quad (2.4)$$

which simplifies to

$$V(PS)=4\left[\sum_{i=1}^{n}p_ir_i^2-\left(\sum_{i=1}^{n}p_ir_i\right)^2\right]. \quad (2.5)$$

For a given set of subjective probabilities, $p_1, p_2, \cdots, p_n$, $V(PS)$ is minimized when $r_i=1/n$ for $i=1, 2, \cdots, n$. If the stated probabilities are all equal to $1/n$, the variance of the probability score is zero, since the score will be the same no matter which of the n outcomes occurs.

Consider the two-state ($n=2$) situation. Then

$$\begin{aligned}E(PS^2)&=1-2p_1r_1-2p_2r_2+r_1^2+r_2^2\\&=1-2p_1r_1-2(1-p_1)(1-r_1)+r_1^2+(1-r_1)^2, \quad (2.6)\end{aligned}$$

and

$$\begin{aligned}V(PS)&=4[p_1r_1^2+p_2r_2^2-(p_1r_1+p_2r_2)^2]\\&=4p_1(1-p_1)(1-2r_1)^2. \quad (2.7)\end{aligned}$$

Note that, for a given p_1, $V(PS)$ is minimized when $r_1=0.5$ and maximized when $r_1=1$ or $r_1=0$.

3. The probability forecasts of a risk-taker

Suppose that the assessor's utility function for positive changes in wealth is quadratic, i.e.,

$$U(x)=x^2, \text{ for } x\geq 0, \quad (3.1)$$

where x represents an increase in the assessor's wealth. This function is convex, and an individual with a convex utility function is a "risk-taker."

Suppose, in addition, that the assessor must make a probability forecast and that he will receive a payoff of $2-PS$ after the event in question is observed. This payoff cannot be negative, since the largest possible value of the probability score is two. Thus, the utility of the payoff to the assessor is

$$\begin{aligned}U(\text{Payoff})&=U(2-PS)=(2-PS)^2\\&=4-4PS+PS^2. \quad (3.2)\end{aligned}$$

The assessor should choose $r_1, r_2, \cdots, r_n$ so as to maximize his expected utility. For any set of the probabilities $r_1, r_2, \cdots, r_n$, the expected utility is

$$\begin{aligned}EU(r_1,\cdots,r_n)&=\sum_{h=1}^{n}p_hU(\text{Payoff})\\&=\sum_{h=1}^{n}p_h(4-4PS_h+PS_h^2)\\&=4-4E(PS)+E(PS^2). \quad (3.3)\end{aligned}$$

Adding and subtracting $[E(PS)]^2$ yields

$$\begin{aligned}EU(r_1, \cdots, r_n)&=4-4E(PS)+E(PS^2)-[E(PS)]^2+[E(PS)]^2\\&=4-E(PS)[4-E(PS)]+V(PS). \quad (3.4)\end{aligned}$$

Thus, the expected utility depends not only on the expected probability score, but also on the variance of the probability score.

In order to investigate the behavior of the optimal probability forecasts, consider the simple two-state situation. Optimal values of r_1 were found for $p_1=0.50$ (0.01) 1.00. Note that, since $r_2=1-r_1$, the determination of the optimal r_1 automatically gives the optimal r_2 as well. Also, because of the symmetry of the two-state situation, the optimal value of r_1 for $p_1<0.50$ can be found by reversing the subscripts, so that p_1 will then be greater than 0.50.

The relationship between p_1 and the optimal r_1 is illustrated in Fig. 1. For $p_1=0$, $p_1=0.50$ or $p_1=1$, $r_1=p_1$, while for all other values of p_1, $r_1\neq p_1$. Note that if $p_1<0.50$, $r_1\leq p_1$, and if $p_1>0.50$, $r_1\geq p_1$. In fact, for $p_1\leq 0.33$, $r_1=0$, and for $p_1\geq 0.67$, $r_1=1$. This result can be explained in terms of the expectation and variance of the probability score, using (3.4). The second term on the right-hand side of (3.4) is maximized when $E(PS)$ is minimized, since the derivative of this term with respect to $E(PS)$ is $-4+2E(PS)$, which is less than or equal to zero because $E(PS)\leq 2$. Thus, the assessor can maximize the second term in (3.4) by minimizing $E(PS)$. Since PS is a proper scoring rule, $E(PS)$ is minimized when $r_1=p_1$. However, consider the third term in (3.4), which is simply $V(PS)$. From (2.7), $V(PS)$ is maximized when $r_1=1$ or $r_1=0$. This situation creates a conflict for the assessor; that is, in order to maximize the second term in (3.4), he should let $r_1=p_1$, while in order to maximize the third term, he should let $r_1=0$ or $r_1=1$.

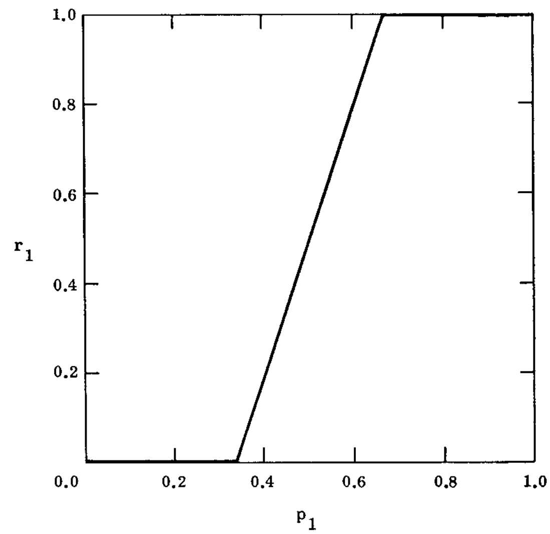

FIG. 1. The optimal forecast r_1 as a function of the subjective probability p_1 for the "risk-taker" in the two-state situation.

Maximizing the entire expression (3.4) requires a combination of these two possibilities. When $p_1<0.50$, the optimal r_1 is between zero and p_1; when $p_1>0.50$, the optimal r_1 is between p_1 and one. When $p_1\leq 0.33$ or $p_1\geq 0.67$, the effect of the third term in (3.4), the variance of the probability score, is such that the second term becomes irrelevant and the optimal forecast is

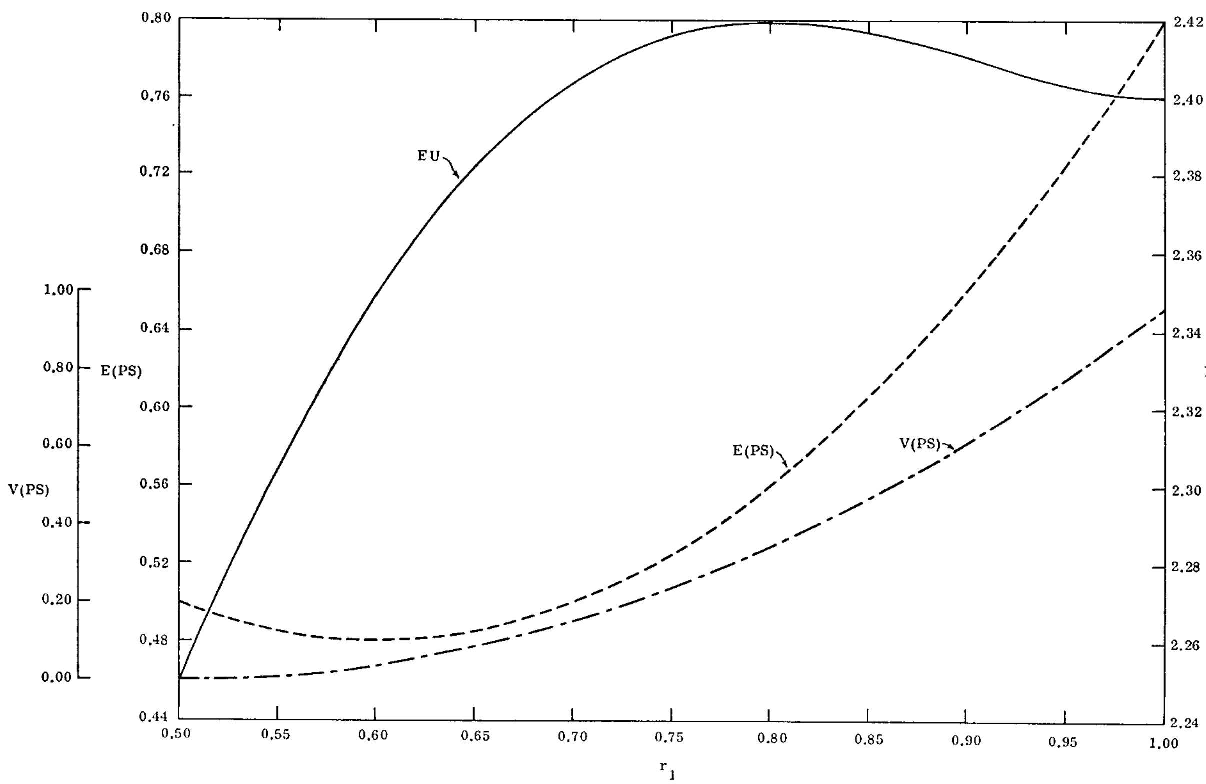

FIG. 2. The expected probability score $E(PS)$, the variance of the probability score $V(PS)$, and the expected utility EU of the forecast r_1 for the "risk-taker" in the two-state situation when $p_1=0.60$.

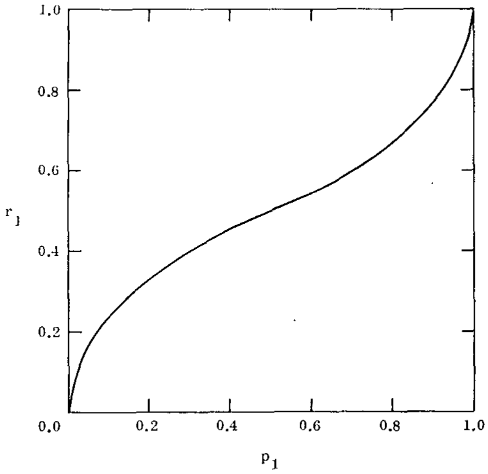

FIG. 3. Same as Fig. 1 except for the "risk-avoider."

categorical. All other things being equal, the assessor prefers a large $V(PS)$ to a small $V(PS)$, which seems reasonable since the assessor's utility function is that of a risk-taker.

As an example of the effect of $E(PS)$ and $V(PS)$ on the optimal probability forecast, suppose that $p_1=0.60$. The three curves in Fig. 2 represent $E(PS)$, $V(PS)$ and EU as a function of r_1. Note that $E(PS)$ is minimized when $r_1=0.60$ and $V(PS)$ is maximized when $r_1=1$. The EU curve is maximized in this case when $r_1=0.80$ (the fact that the optimal value of r_1 is exactly midway between p_1 and one is simply a coincidence).

We should mention that, for some utility functions, the optimal probability forecast would always be either zero or one. For instance, consider the utility function of an *extreme* risk-taker (for the probability assessment task):

$$U(x)=\begin{cases}0, & \text{if} \quad 0\leq x<2\\ 1, & \text{if} \quad x\geq 2\end{cases} \tag{3.5}$$

For this utility function, the optimal forecast is $r_1=1$ if $p_1\geq 0.50$ and $r_1=0$ if $p_1\leq 0.50$. This is obviously a pathological case, but surely other (nonpathological) utility functions exist which give the same results. In the n-state case, the optimal forecast for the extreme risk-taker is to set one r_i (the one corresponding to the largest p_i) equal to one and the remaining r_i equal to zero.

4. The probability forecasts of a risk-avoider

Suppose that the assessor's utility function for positive changes in wealth is exponential, i.e.,

$$U(x)=1-e^{-x}, \text{ for } x\geq 0, \tag{4.1}$$

where x represents an increase in the assessor's wealth. This function is concave, and an individual with a concave utility function is a "risk-avoider."

As in the previous section, suppose, in addition, that the assessor must make a probability forecast and that he will receive a payoff of $2-PS$ after the event in question is observed. The utility of the payoff to the assessor is

$$U(\text{Payoff})=U(2-PS)=1-e^{-(2-PS)}. \tag{4.2}$$

The expected utility for any set of probabilities $r_1, r_2, \cdots, r_n$ is

$$\begin{aligned}EU(r_1,\cdots,r_n)&=\sum_{h=1}^{n} p_h U(\text{Payoff})\\ &=\sum_{h=1}^{n} p_h[1-e^{-(2-PS_h)}]\\ &=1-E(e^{PS-2}). \end{aligned} \tag{4.3}$$

In order to investigate the behavior of the optimal probability forecasts, the simple two-state situation was considered, as in the previous section. The relationship between p_1 and the optimal r_1 is illustrated in Fig. 3. For $p_1=0$, $p_1=0.5$ or $p_1=1$, $r_1=p_1$, while for all other values of p_1, $r_1\neq p_1$. Note that if $p_1<0.50$, $p_1\leq r_1\leq 0.5$, and if $p_1>0.50$, $0.5\leq r_1\leq p_1$.

Although in this case we cannot express EU as a simple function of $E(PS)$ and $V(PS)$, we can reasonably assume that a risk-avoider should prefer a small variance to a large variance. This contention is supported to some extent by Fig. 4, which refers to the situation in which $p_1=0.60$. The $E(PS)$ and $V(PS)$ curves are identical to those in Fig. 2, but the EU curve is quite different because of the different utility function. The risk-avoider, instead of "hedging" toward the end points of the unit interval, "hedges" toward the midpoint, $p_1=0.50$.

For some utility functions, the optimal probability forecast in the two-state situation would always be 0.5. For instance, consider the following utility function of an *extreme* risk-avoider (for the probability assessment task):

$$U(x)=\begin{cases}0, & \text{if} \quad 0\leq x<\frac{3}{2}\\ 1, & \text{if} \quad x\geq \frac{3}{2}\end{cases} \tag{4.4}$$

Since the assessor can assure himself of a score of 0.5 and a payoff of $2-PS=1.5$ by assessing $r_1=r_2=0.5$, he has no reason to give a different assessment and risk a lower payoff. As was the case with (3.5), (4.4) obviously represents a pathological case, but surely other (nonpathological) utility functions exist which give the same results. In the n-state case, the optimal forecast for the extreme risk-avoider is $r_i=1/n$ for $i=1, 2, \cdots, n$.

5. Nonlinear utility and probability assessment

What are the implications of nonlinear utility functions for the process of probability assessment (with re-

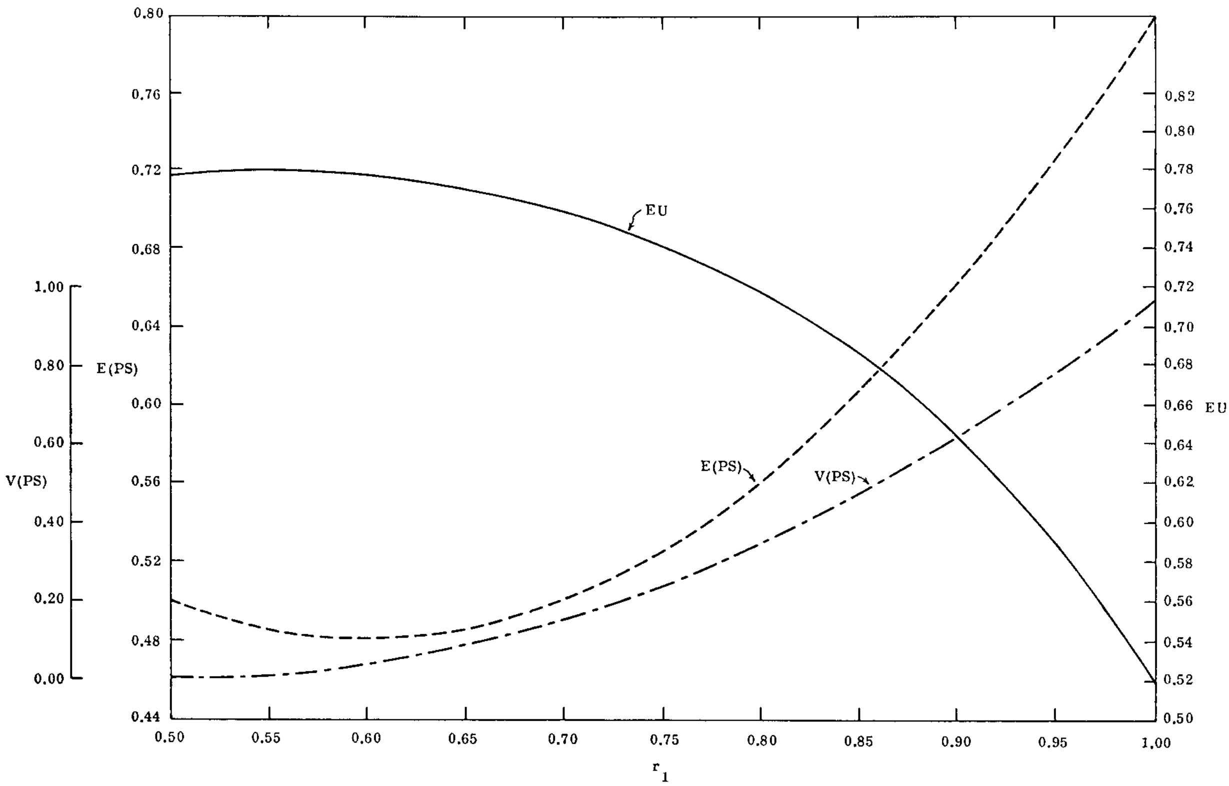

FIG. 4. Same as Fig. 2 except for the "risk-avoider".

gard to the utilization of scoring rules)? The answer to this question depends, in part, upon the nature of the assessor's knowledge concerning his utility function.

If the assessor is able to specify his utility function, then this function can, and should, be incorporated into the assessment process. This is accomplished by defining a new scoring rule, a composite function of the original scoring rule and the (nonlinear) utility function, which is proper under this utility function (Winkler, 1969). Then, the assessor maximizes (or minimizes) his expected score *and* maximizes his expected utility by setting r_i equal to p_i for all i. Thus, in this situation, the assessor should not utilize the original scoring rule and "hedge," but should determine the new (composite) scoring rule and make his statements correspond to his judgments.

If, on the other hand, the assessor is unable to specify his utility function, then the function cannot, of course, be incorporated into the assessment process. Therefore, the assessor's statements may differ from his judgments. What can we say about the nature of these differences? We described in Sections 3 and 4 the differences which would result if the assessor's utility function assumed certain specific forms. The results indicate that for some utility functions the differences would be large. However, for other, approximately linear, utility functions the differences would, no doubt, be small.[6] From a practical (meteorological) point of view, on the other hand, very little is known about these differences.

The basic problem, then, is the determination of the assessor's utility function. One approach is to determine the assessor's utility function through the process of interrogation, i.e., by asking the assessor about his preferences. Another approach is to attempt to determine the assessor's utility function through an analysis of his past behavior in similar situations. We plan to utilize this latter approach, i.e., the analysis of the meteorologist's forecasts to "estimate" his utility function, in a future study.

6. Summary and conclusion

In this paper we demonstrated that, for nonlinear utility functions, factors other than the expected score may be relevant in determining a probability forecast. A case was presented in which the variance of the score was shown to be an important factor, and other utility functions could probably be found for which higher

[6] The judgments and the statements (forecasts) are assumed to be the products of serious deliberations on the part of the assessor. The assessor's deliberations relate to both his beliefs (judgments) and his preferences (utilities).

order moments of the probability score would be of interest.

Whenever a payoff (either explicit or implicit) is associated with a scoring rule such as the probability score, the utility function of the assessor may cause him to "hedge" in some manner. Generally, a risk-taker will "hedge" toward a categorical forecast, and a risk-avoider will "hedge" away from a categorical forecast (i.e., toward a forecast in which all the probabilities are equal). This situation is a universal problem in decision theory, for most formulations of decision-making problems have assumed that the decision maker's utility function is linear with respect to money. If the form of a nonlinear utility function is known, we can, as indicated in Section 5, include this function in the analysis. The main difficulty lies in the determination of an individual's utility function, and more work is needed on this vital problem.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Fishburn, P. C., 1964: *Decision and Value Theory.* New York, Wiley, 451 pp.

Murphy, A. H., 1969: The probability score. Hartford, Conn., Travelers Research Corporation, manuscript, 52 pp.

——, 1970: To "hedge" or not to "hedge": A note on "hedging" and the probability score. *Mon. Wea. Rev.*, **98** (in press.)

Savage, L. J., 1954: *The Foundations of Statistics.* New York, Wiley, 294 pp.

Winkler, R. L., 1969: Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.*, **65**, 1073–1078.

——, and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.*, **7**, 751–758.